ISSN 1392-8619 print/ISSN 1822-3613 online
ŪKIO TECHNOLOGINIS IR EKONOMINIS VYSTYMAS
TECHNOLOGICAL AND ECONOMIC DEVELOPMENT OF ECONOMY
http://www.tede.vgtu.lt
2006, Vol XII, No 4, 289–294

TESTING THE RELATIONAL PERSPECTIVE MAP FOR VISUALIZATION OF MULTIDIMENSIONAL DATA

Rasa Karbauskaitė [1,2], Virginijus Marcinkevičius[1], Gintautas Dzemyda[1,2]

[1]Institute of Mathematics and Informatics, Akademijos g. 4, LT-08663 Vilnius, Lithuania
[2]Vilnius Pedagogical University, Studentų g. 39, LT-08106 Vilnius, Lithuania
E-mails: Karbauskaite@ktl.mii.lt, VirgisM@ktl.mii.lt, Dzemyda@ktl.mii.lt

Received 15 June 2006; accepted 20 November 2006

Abstract. This paper deals with a method, called the relational perspective map that visualizes multidimensional data onto two-dimensional closed plane. It tries to preserve the distances between the multidimensional data in the lower-dimensional space. But the most important feature of the relational perspective map is the ability to visualize data in a non-overlapping manner so that it reveals small distances better than other known visualization methods. In this paper, the features of this method are explored experimentally and some disadvantages are noticed. We have proposed a modification of this method, which enables us to avoid them.

Keywords: visualization, multidimensional data, relational perspective map.

1. Introduction

Data perception is frequently a complex problem, especially when data arise from a complicated phenomenon described by many parameters, i.e., multidimensional data are analyzed. In order to better perceive multidimensional data, to establish their interrelations, and the groups (clusters) formed, we often have to visualize them. A human being is capable to perceive visual information much faster than textual.

Visualization methods of multidimensional data can be partitioned into several groups: direct visualization methods, projection methods, clustering methods, and methods based on artificial neural networks. Projection methods are frequently used, the aim of which is to present multidimensional data in a space of smaller dimension so as to preserve the analyzed data structure as precisely as possible. There are linear projection methods (Principal Component Analysis (PCA) [1], Projection Pursuit [2], etc.) and nonlinear projection methods (Multidimensional Scaling (MDS) [3], Sammon's algorithm [4], Principal Curves [5], the Triangulation method [6], the relational perspective map (RPM) [7], etc.).

In this paper, we investigate the relational perspective map (RPM) method [7]. It visualizes multidimensional data onto the closed plane (torus surface) so that the distances between data in the lower-dimensional space would be as close as possible to the original distances. But what is more important, the RPM method also gives the ability to visualize data in a non-overlapping manner so that it reveals small distances better than other known visualization methods. It is shown in [7].

In this paper, we also propose a modification of the RPM method and experimentally explore the features of the RPM method and its modification. The main feature of this modification is its independence on a size of torus.

2. The RPM method

In this section, we present some essential details on the RPM method from [7]. Assume, we have a set of data points $S = \{s_i = (s_{1i},...,s_{ni}), i = 1,...,N\}$, $s_i \in R^n$ with a distance matrix $\delta_{ij}, i, j = 1,...,N$; N is the number of data points. The RPM algorithm maps data points s_i into image points t_i in a two-dimensional space (torus surface) in such a way that visual distances between the image points, denoted by $d_{ij}, i, j = 1,...,N,$ resemble the distances δ_{ij}. We call δ_{ij} and d_{ij}, respectively, relational and image distance matrices.

As illustrated in Fig 1, the RPM algorithm first maps data points onto the surface of a torus, then onto the flat rectangle by a vertical and a horizontal cut.

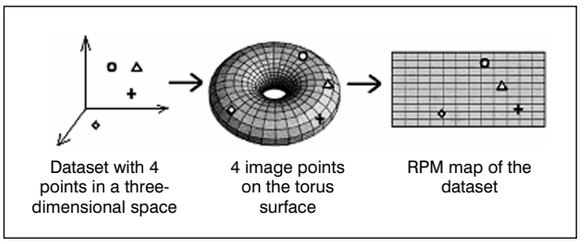


Fig 1. The model of the RPM method

From the physical point of view, the torus is a force-directed multiparticle system: the image points are considered as particles that can move freely on the surface of the torus, but cannot escape the surface. The particles exert repulsive forces on one another so that, guided by the forces, the particles rearrange themselves to a configuration that visualizes the relational distances δ_{ij}.

The RPM algorithm uses equation (1) as the total potential energy to characterize a configuration.

$$E_p = \sum_{i<j} \frac{\delta_{ij}}{p d_{ij}^p}, \; with \; E_0 = -\sum_{i<j} \delta_{ij} \ln\left(d_{ij}\right), \qquad (1)$$

$p \in (-1.0; +\infty)$. The forces between the particles are characterized by:

$$f_{ij} = \frac{\partial E_p}{\partial d_{ij}} = -\frac{\delta_{ij}}{d_{ij}^{p+1}}, \;\; \text{where } i < j. \qquad (2)$$

In order to derive a practical algorithm, a more formal specification of the torus and this visualization is given. Let $T := [0, w] \times [0, h] \subset R^2$ denote the rectangle plane of *w* and height *h* in the 2-D Cartesian coordinator system. A torus mapping is understood as a visualization of the following form:

$$\varphi : S \to T, \; s_i \to t_i := (x_i, y_i). \qquad (3)$$

Let $t_i := (x_i, y_i)$, $t_j := (x_j, y_j)$ be two points from *T*. Then the distance between t_i and t_j is defined as follows:

$$d(t_i, t_j) = \min\left\{ |x_i - x_j|, w - |x_i - x_j| \right\} + \min\left\{ |y_i - y_j|, h - |y_i - y_j| \right\}. \qquad (4)$$

With this distance function, the opposite edges of the rectangle *T* are actually stuck together, so that it becomes topologically equivalent to a torus (Fig 1).

So the torus surface used in the RPM algorithm is actually the metric space (T, d). The goal of the RPM algorithm is thus to find a torus mapping φ of form (3) that minimizes equation (1).

The RPM algorithm adapts the Newton-Raphson (NR) [8] method to minimize the energy function (1). Let $f(x)$ be a single-variate function, then the optimum point of $f(x)$ can be found by the following formula:

$$x^{(m+1)} = x^{(m)} - \frac{f'\left(x^{(m)}\right)}{f''\left(x^{(m)}\right)}. \qquad (5)$$

In order to apply the NR method, it is necessary to calculate the first-order and second-order partial derivatives of *E* with respect to all variables x_i and y_i. There will be given formulas of derivatives with respect to x_i; the calculation with respect to y_i is completely analogous.

The first-order partial derivative of *E* is calculated as follows:

$$\frac{\partial E_p}{\partial x_i} = \sum_{i<k} h_{ik} f_{ik}, \quad i, k = 1, ..., N, \qquad (6)$$

f_{ik} is defined in equation (2) and h_{ik} can be calculated as follows:

$$h_{ik} = \frac{\partial d_{ik}}{\partial x_i} = \begin{cases} +1, \; if \; |x_i - x_k| < \dfrac{w}{2}, \;\; x_i > x_k, \\ -1, \; if \; |x_i - x_k| < \dfrac{w}{2}, \;\; x_i < x_k, \\ -1, \; if \; |x_i - x_k| > \dfrac{w}{2}, \;\; x_i > x_k, \\ +1, \; if \; |x_i - x_k| > \dfrac{w}{2}, \;\; x_i < x_k. \end{cases} \qquad (7)$$

If $|x_i - x_k| = \frac{w}{2}$ or $|x_i - x_k| = 0$, then $h_{ik} = 1$ or $h_{ik} = -1$ may be used. We have chosen $h_{ik} = 1$ in the realization of RPM algorithm. The second-order partial derivative of *E* is calculated as follows:

$$\frac{\partial^2 E_p}{\partial x_i^2} = -(p+1) \sum_{i<k} \frac{f_{ik}}{d_{ik}}. \qquad (8)$$

By substituting (6) and (8) into formula (5), we get an iterative formula to find the minimum energy configuration:

$$x_i^{(m+1)} = x_i^{(m)} + \frac{1}{p+1} \frac{\sum_{i<k} h_{ik} f_{ik}}{\sum_{i<k} \frac{f_{ik}}{d_{ik}}}. \qquad (9)$$

James Xinzhi Li [7] uses a modified variant of formula (9):

$$x_i^{(m+1)} = x_i^{(m)} + c^{(m)} \frac{\sum_{i<k} h_{ik} f_{ik}}{\sum_{i<k} \frac{f_{ik}}{d_{ik}}}. \qquad (10)$$

In formula (10), the constant $\frac{1}{p+1}$ is replaced by the

parameter $c^{(m)}$ that is called the learning speed at the step m. Here $c^{(m)}$ should approach zero as m increases. $c^{(m)}$ is calculated by the formula:

$$c^{(m)} = ra^m, \tag{11}$$

where r is the initial learning speed, $a \in (0;1)$. Both r and a are determined empirically.

3. Results of experimental investigation of the RPM method

The following datasets were used in the experiments:

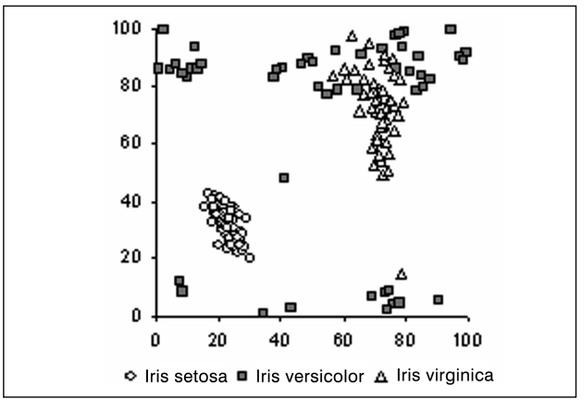
- **The spherical dataset.** It is a set of 576 points, which components (x, y, z) are calculated according to the parametrical equations below:

$$\begin{cases} x = 2\cos\alpha\cos\beta, \\ y = 2\sin\alpha\cos\beta, \\ z = 2\sin\beta, \end{cases} \tag{12}$$

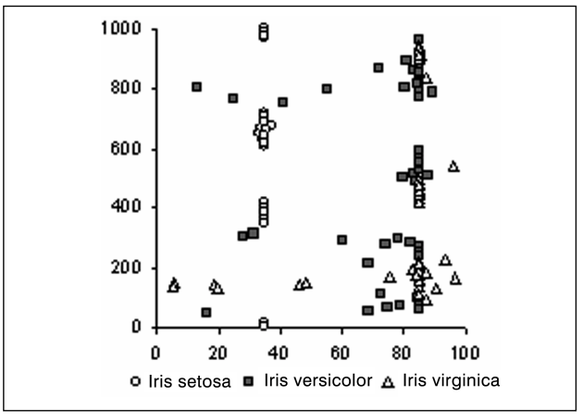
  varying the values of the parameters α and β at equal intervals, where $\alpha \in [0;360^\circ]$, $\beta \in [0;360^\circ]$.
- **The classical Fisher iris dataset** [9]. Petal weights, petal heights, sepal weights, and sepal heights of 150 iris flowers were measured using 50 flowers of three different kinds: Iris Setosa, Iris Versicolor, and Iris Virginica. The dataset consists of 150 four-dimensional points.
- **The HBK dataset** [10]. 75 four-dimensional points comprised three separate groups: 1–10 points form the first group, 11–14 points – the sec-ond one, and 15–75 points – the third group.

In [7], it is not indicated exactly, what values of the parameters w (the width of the rectangle plane), h (the height of the rectangle plane), and r (11) should be in order to minimize the potential energy and obtain a visualization of multidimensional points as precisely as possible.

We have done many experiments with three datasets – sphere, iris, and HBK – to find out how much the obtained visualizations of multidimensional points depend on the values of the parameters r, w, h (Fig 2, 3). In [7], it was proposed that a should be near to 1, and we have chosen $a = 0.975$ in our experiments. We have chosen $p = 0$ as recommended in [7]. The initial coordinates of points on the plane are generated at random.

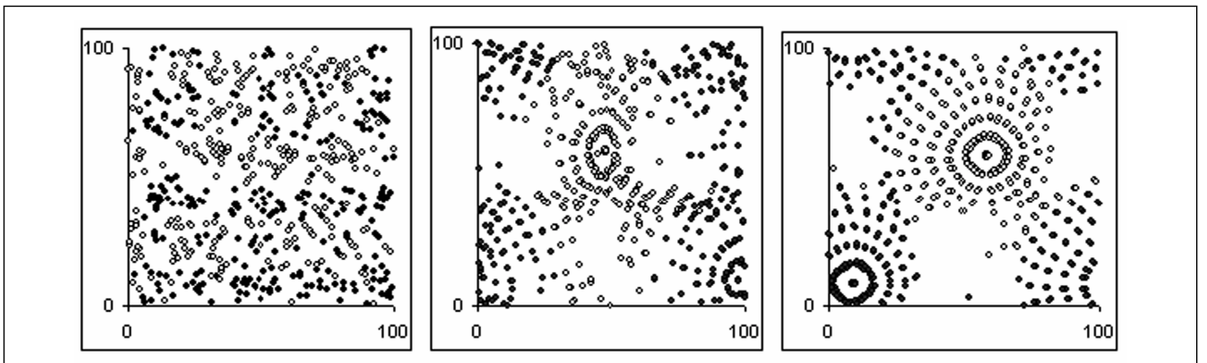


a) $w = 100$, $h = 100$



b) $w = 100$, $h = 1000$

Fig 2. Visualization of the iris dataset on the plane by the RPM method at different values of the parameters w and h



$r = 0.5$, $E = -3342223{,}523$ $r = 2$, $E = -3472738.457$ $r = 4$, $E = -3491417.189$

Fig 3. Visualization of the spherical dataset on the plane by the RPM method at different values of the parameter r

When visualizing the iris dataset, different values of the parameters w and h were chosen, and the change of the potential energy was explored. The values of the parameters a and r were fixed: $a = 0.975$, $r = 4$. 10 experiments have been done for every couple of parameters w and h (for example, $w = 100$, $h = 100$) and the obtained values of the potential energy have been averaged. It has been observed that the larger the ratio of these parameters, the less potential energy is obtained (Fig 4), however, the visualization quality of data points is worse: the classes are not formed (Fig 2 a, b). Thus, having rectangle planes of different width and height, we cannot compare the minima of the potential energy and state that the less potential energy, the more precise visualization of the data points. Analogical results are obtained when visualizing the HBK dataset.

When pursuing investigations with the spherical dataset, we have also kept a close watch on the dependence of the potential energy on the parameter r. The values of the parameters w, h, a were fixed: $w = h = 100$, $a = 0.975$ Fig 5 illustrates, how the potential energy is changing at different values of the parameter r when visualizing the dataset of sphere. Though the relative value of the potential energy does not change considerably (at maximum it changes between $E_p(r = 0.5)$ and $E_p(r = 4)$ in 4 %), when changing r, however, even an insignificant reduction of the potential energy substantially changes the projections of points (Fig 3).

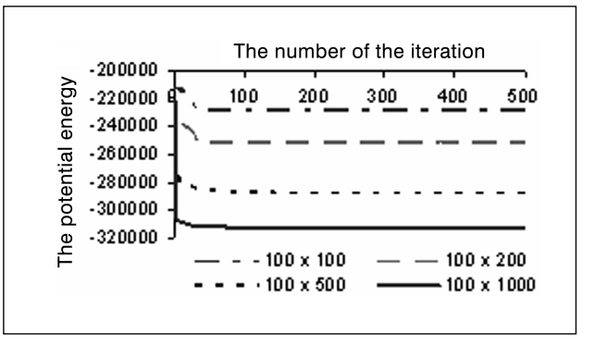


Fig 4. Dependence of the potential energy on the parameters w and h while visualizing iris data by the RPM method

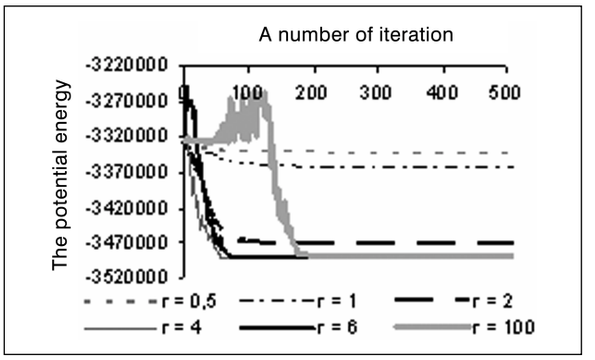


Fig 5. Dependence of the potential energy on the parameter r, while visualizing the spherical dataset by the RPM method

4. A modification of the RPM method

In this paper, we have proposed a modification of the RPM method. The main ideas are as follows:

- the distance function on T is defined in another way;
- in each iteration we recalculate the projections not of all the points at once, but we pick a point one by one and recalculate its coordinates, taking into consideration the points whose coordinates have already been recalculated and those whose coordinates have not been changed as yet.

The distance function on T is defined in such a way:

$$d\left(t_i, t_j\right) = \min\left\{\frac{\left|x_i - x_j\right|}{w}, 1 - \frac{\left|x_i - x_j\right|}{w}\right\} + \min\left\{\frac{\left|y_i - y_j\right|}{h}, 1 - \frac{\left|y_i - y_j\right|}{h}\right\}. \qquad (13)$$

Having introduced such a distance function, we obtain that all $d\left(t_i, t_j\right) \in [0;1]$. Because of the distance function (13), we have to recalculate formulas (7), (8), and (9):

$$h_{ik} = \frac{\partial d_{ik}}{\partial x_i} = \begin{cases} +\frac{1}{w}, \text{ if } \left|x_i - x_k\right| < \frac{w}{2}, \ x_i > x_k, \\ -\frac{1}{w}, \text{ if } \left|x_i - x_k\right| < \frac{w}{2}, \ x_i < x_k, \\ -\frac{1}{w}, \text{ if } \left|x_i - x_k\right| > \frac{w}{2}, \ x_i > x_k, \\ +\frac{1}{w}, \text{ if } \left|x_i - x_k\right| > \frac{w}{2}, \ x_i < x_k. \end{cases} \qquad (14)$$

If $\left|x_i - x_k\right| = \frac{w}{2}$ or $\left|x_i - x_k\right| = 0$, then $h_{ik} = \frac{1}{w}$ or $h_{ik} = -\frac{1}{w}$ may be used. We have chosen $h_{ik} = \frac{1}{w}$.

The second-order partial derivative of E:

$$\frac{\partial^2 E_p}{\partial x_i^2} = -\frac{p+1}{w^2}\sum_{i<k}\frac{f_{ik}}{d_{ik}}. \qquad (15)$$

An iterative formula to find the minimum energy configuration:

$$x_i^{(m+1)} = x_i^{(m)} + \frac{w^2}{p+1}\frac{\sum_{i<k} h_{ik} f_{ik}}{\sum_{i<k}\frac{f_{ik}}{d_{ik}}} + wK. \qquad (16)$$

In (16), the product wK is added to satisfy the condition $0 \le x_i^{(m+1)} < w$, where K is the integer number, selected to meet this rule.

The calculation with respect to y_i is completely analogous.

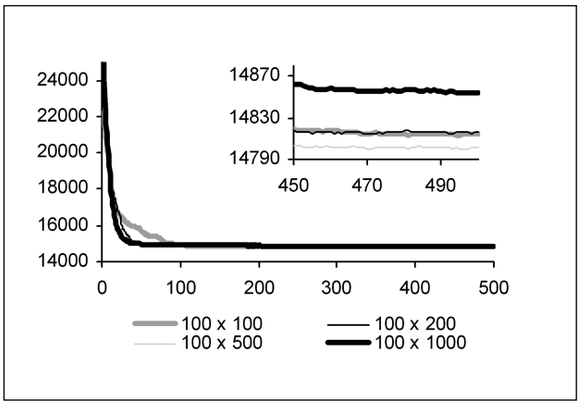
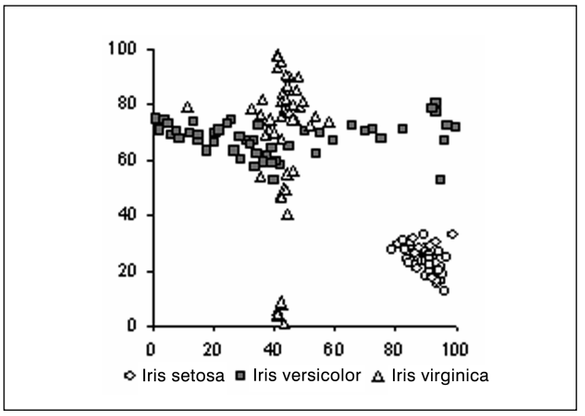
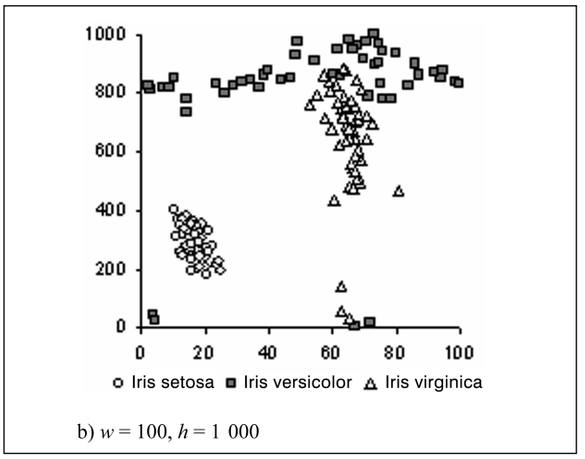


Fig 6. Dependence of the potential energy on the parameters w and h while visualizing iris data by the modification of the RPM method



a) $w = 100$, $h = 100$



b) $w = 100$, $h = 1\ 000$

Fig 7. Visualization of the iris dataset on the plane by the modification of the RPM method

5. Results of experimental investigation of the modification of the RPM method

When visualizing the iris dataset, different values of the parameters w and h were chosen and the change of the potential energy was explored. 10 experiments have been done for every couple of the parameters w and h (for example, w = 100, h = 100) and the values of the potential energy obtained have been averaged. It has been noticed that, in the case of the RPM method modification, we avoid a strong dependence of the energy on the width and height of the rectangle plane (Fig 6) and the visualization of data points is similar at different parameters w and h (Fig 7). In the case of our modification, the relative value of the potential energy changes at maximum by 0.4 % (Fig 6), while in the case of the RPM method, it may change even up to 27 % (Fig 4). Note: the potential energy does not gradually converge to the minimum (Fig 6).

Analogical results have been obtained when visualizing the HBK dataset. Fig 8 illustrates the visualization of the spherical dataset when $w = h = 100$.

6. Conclusions and discussion

Experiments with various datasets have shown that algorithms of the RPM type may be helpful in visualizing multidimensional data. The results of the basic RPM algorithm strongly depend on the parameters w, h, a and r. Unfortunately, there are no specific rules to select the values of these parameters. In our modification, we avoid a strong dependence on the parameters w, h. However, having rejected the parameters a and r, the energy does not gradually converge to the minimum. When pursuing investigations, it has been noticed that the optimization process stabilizes after 100 iterations. In order to define the stopping condition of the algorithm, additional detailed investigations are necessary.

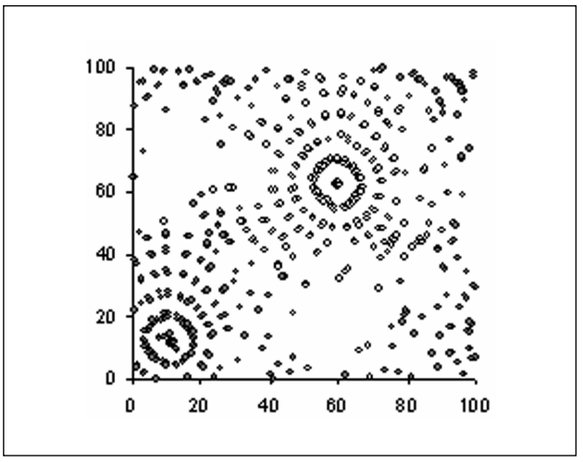


Fig 8. Visualization of the spherical dataset on the plane by the modification of the RPM method

The most important drawback of the RPM method is that the function E_p is non differentiable in some points: if there are pairs x_i and x_j, for which $x_i = x_j$ or $|x_i - x_j| = w/2$, then, in the close space of the point x_i, the values of the right and the left derivatives of the function do not coincide. Applying the NR method, we choose the value of the left derivative of the function. The better approach is to use search methods, which do not assume differentiability, or the distance function on T would be defined as the function differentiable in all points of the torus surface.

Acknowledgements

The authors thank the anonymous reviewer for valuable remarks, which improved the quality of this paper and gave ideas for future investigations.

References

1. Taylor, P. Statistical Methods. Intelligent Data Analysis: an Introduction, edited by M. Berthold, D. J. Hand. Springer-Verlag, 2003, p. 69–129.
2. Brunsdon, C.; Fotheringham, A. S.; Charlton, M.E. An investigation of methods for visualising highly multivariate datasets in case studies of visualization in the social sciences, D. Unwin and P. Fisher (eds.) Joint Information Systems Committee, ESRC, Technical Report Series 43, p. 55–80, 1998. ISSN 1356-9066.
3. Borg, P. J. F. Groenen. Modern multidimensional scaling. 2nd edition. New York: Springer, 2005.
4. Sammon, J. W. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, Vol 18, 1969, p. 401–409.
5. Hastie, T. Principal Curves and Surfaces. PhD Dissertation. Stanford Linear Accelerator Center, Stanford University, Stanford, California, 1984. http://www.slac.stanford.edu/pubs/slacreports/slac-r-276.html
6. Lee, R. C. T.; Slagle, J. R.; Blum, H. A triangulation method for the sequential visualization of points from N-space to two-space. *IEEE Transactions on Computers*, Vol 26, 1977, p. 288–292.
7. James Xinzhi Li. Visualization of High Dimensional Data with Relational Perspective Map. *Information Visualization*, Vol 3, No 1, 2004, p. 49–59.
8. Press, W. H.; Teukolsky, S. A.; Vetterling, W. T. Flannery BP. Numerical Recipes in C++ (2nd Edition). Cambridge University Press: Cambridge, 2002, p. 366–372.
9. Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, Vol 7, 1936, p. 179–188,.
10. Hawkins, D. M.; Bradu, D.; Kass, G. V. Location of several outliers in multiple regression data using elemental sets. *Technometrics*, 26, 1984, p. 197–208.

SANTYKINĖS PERSPEKTYVOS METODO, SKIRTO DAUGIAMAČIAMS DUOMENIMS VIZUALIZUOTI, TYRIMAS

R. Karbauskaitė, V. Marcinkevičius, G. Dzemyda

Santrauka

Nagrinėjamas santykinės perspektyvos metodas (angl. *relational perspective map* (RPM)), kuris vizualizuoja daugiamačius duomenis į plokštumą. RPM metodas kaip ir dauguma kitų žinomų dimensijos mažinimo metodų stengiasi išlaikyti santykinius atstumus tarp daugiamačių taškų plokštumoje. Pagrindinė RPM metodo savybė ta, kad duomenys vizualizuojami plokštumoje taip, kad jų projekcijos nepersidengtų. Taigi RPM metodas išlaiko atstumus tarp artimų taškų daug tiksliau negu kiti vizualizavimo metodai. Eksperimentais ištyrus RPM metodą, nustatyti šio metodo trūkumai, todėl pasiūlyta modifikacija, leidžianti jų išvengti.

Reikšminiai žodžiai: vizualizavimas, daugiamačiai duomenys, santykinės perspektyvos metodas.

Rasa KARBAUSKAITĖ. Doctoral student. Engineer programmer, Department of Systems Analysis, Institute of Informatics and Mathematics. Junior lecturer, Department of Informatics, Vilnius Pedagogical University. Bachelor's degree in Mathematics and Informatics (2003), Master's degree in Informatics (2005) from Vilnius Pedagogical University. Research interests: visualization of multidimensional data, clustering.

Virginijus MARCINKEVIČIUS. Doctoral student. Engineer programmer, Department of Systems Analysis, Institute of Informatics and Mathematics. Bachelor's degree in Mathematics and Informatics (2001), Master's degree in Mathematics (2003) from Vilnius Pedagogical University. Research interests: neural networks, data mining, parallel computing.

Gintautas DZEMYDA. Professor, Doctor Habil. Department of Informatics, Vilnius Pedagogical University. Department of Systems Analysis, Institute of Informatics and Mathematics (IMI). Doctor's degree in Technical Sciences in 1984 after post-graduate studies at IMI. Doctor Habil from Kaunas University of Technology in 1997. Research interests: interaction of optimisation and data analysis, optimisation theory and applications, multiple criteria decisions, neural networks, and data analysis.